



Investigating word interactions in texts. Application to text categorization in genomics

Martine Cadot, Michel Zitt, Gabriel Meurin, Alain Lelu

► To cite this version:

Martine Cadot, Michel Zitt, Gabriel Meurin, Alain Lelu. Investigating word interactions in texts. Application to text categorization in genomics. First SaarLorLux Workshop on Systems Biology 2009, Computational, Structural and Medical Approaches for Systems Biology, Dec 2009, Nancy, France. inria-00442395

HAL Id: inria-00442395

<https://inria.hal.science/inria-00442395>

Submitted on 21 Dec 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Investigating word interactions in texts. Application to text categorization in genomics.

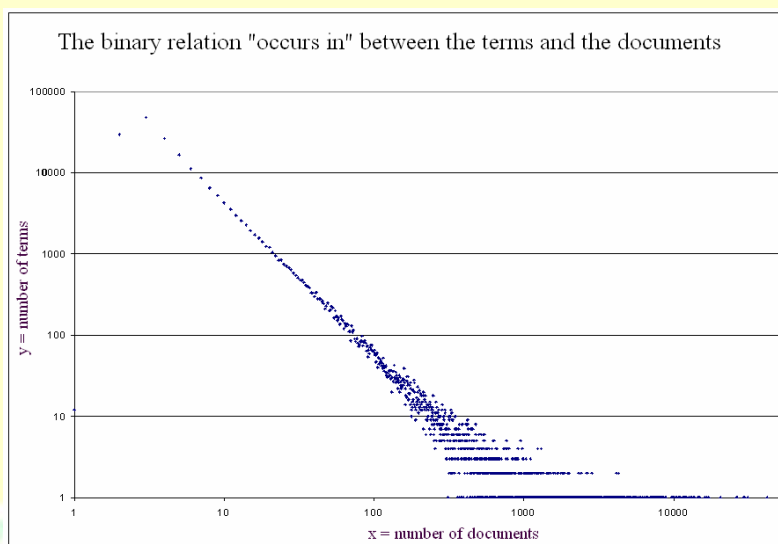
Martine Cadot (LORIA), Michel Zitt (INRA, OST), Gabriel Meurin (LORIA/INRA), Alain Lelu (LORIA, LASELDI)

Words interacting in a text may be compared, to a certain extent, to molecules interacting and building “complexes”, i.e. phrases, named entities, or longer-range semantic or syntactic associations. We will call them “k-itemsets”, k being their interaction level. We have shown (Cadot 06) that an adequately built subset of these k-itemsets is enough for describing the entirety of the relations at work in a corpus represented as a set of “bag-of-words” documents, whatever the level k of these relations.

Our objective is to reconstruct each category into which a corpus of scientific abstracts has been split, using a set of Boolean queries as a best compromise between conciseness and reproducibility of this categorization.

I - The corpus, its 50 catégories.

- Scientific abstracts in genomics, pulled out from the Web of Science (Thomson Scientific ed.).
- This subset has been delineated using a hybrid method, based both on lexical queries and citation expansion/ shrinkage (Zitt et al. 2006) → 120,000 abstracts from 1999 to 2005.
- A vocabulary of 237,000 lemmatized words and phrases (>2 occurrences) has been pulled out and filtered (NeuroNav, www.diatopie.com). I.e.: *sequence, polymorphism, folded_structure, chromosome_4B, greenbug_resistance_gene,...*



This figure reads: e.g. 1047 terms occur each one exactly in 21 documents.

- 50 categories resulted from a clustering of the abstracts by the Axial K-means method (Bassecoulard et al. 2007).

M 1/ Human_genome/ Human_genome_project	M 17/ Map/ Linkage_maps/ Polymorphism	M 33/ RNA- Virus
M 2/ Translocation/ FISH/ leukemia	M 18/ Population_genomics	M 34/ PCR/ Methods/ applications
M 3/ Plant_genomics/ Transgenic_plants	M 19/ Repair/ DNA_damage	M 35/ C-DNA/ Transcription/ C-DNA_library
M 4/ DNA_sequence/ Satellite	M 20/ Resistance/ Resistance_genes/ Plant & Trout_resistance	M 36/ Polymorphism
M 5/ Strain/ Microbial_genomics	M 21/ Hybrid/ Somatic_hybrids/ Fertility	M 37/ Cell/ DNA_damage
M 6/ Cell_identity & Gene_expression	M 22/ Human/ C-DNA/ Gene_annotation	M 38/ Genome/ Genome_sizes
M 7/ Enzyme/ Escherichia_Coli	M 23/ Exon/ Genomic_organization/ Gene_annotation	M 39/ DNA/ Arrays/ Genomic_techniques
M 8/ Alignment/ Bioinformatics	M 24/ System/ Systems_biology/ Bioinformatics	M 40/ QTL/ Trait/ Mapping/ Polymorphism
M 9/ Genome/	M 25/ Patient/ Disease_genomics/ Biomarkers/ Pharmacogenomics	M 41/ Signaling/ Kinase/ MAPK
M 10/ Comparative_genomic_hybridization/ Tumor	M 26/ Virus/ Nucleotide_sequence	M 42/ Mutation/ Missense_mutation
M 11/ SNPs/ Polymorphism	M 27/ Evolution/ Evolutionary_genomics	M 43/ Mouse/ Murine_genomics
M 12/ Network/ Biological_networks/ Model	M 28/ Cancer/ Genome & cancer	M 44/ Expression/ Cell_identity & Gene_expression
M 13/ Transcriptional/ Saccharomyces_cerevisiae/ Transcriptome	M 29/ Promoter/ Transcription	M 45/ LOD/ Linkage_analysis/ Polymorphism
M 14/ Locus/ Microsatellite_locus/ Polymorphism	M 30/ Mutant/ Mutagenesis	M 46/ Human/ Primate/ Gene_annotation/ Comparative_genomics
M 15/ Cell_line/ Tumor/ Genome & Cancer	M 31/ LOH/ Tumor_suppressor/ Genome & Cancer	M 47/ Species/ Phylogeny/ Evolutionary_genomics
M 16/ Spectrometry/ Proteomics	M 32/ Marker/ RAPD/ AFLP/ Polymorphism	M 48/ C57BL/ Congenic_strains/ Murine_genomics
		M 49/ Residue/ Amino_acid_sequence
		M 50/ Virus/ Virus_replication/ Virus_recombination

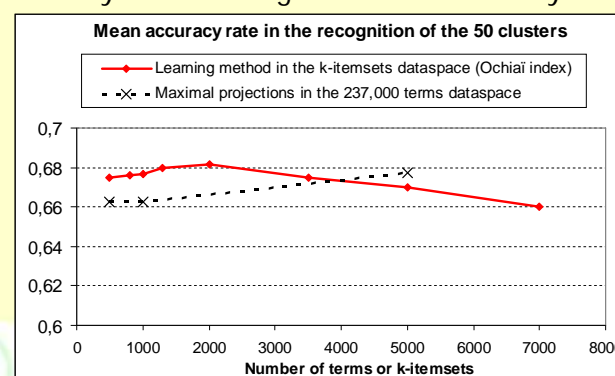
Ex. of the 14 first terms (words and phrases) most typical of the « Human Genome Project » cluster:

No	Phrase	Ochiai			
1	human_genome	0,467	8	human_genome_sequence	0,122
2	human	0,336	9	primate	0,107
3	genome	0,257	10	sequence	0,103
4	human_genome_project	0,238	11	chimpanzee	0,099
5	project	0,157	12	completion	0,096
6	draft	0,138	13	disease	0,094
7	human_chromosome	0,125	14	genomics	0,092
			15	human_gene	0,092

As most of data analysis methods do, this data partition takes into account the only « 2-itemsets » (a k-itemset of support s is an elementary association of k terms present in s documents).

II - Concise and reproducible representation of the 50 catégories: using itemsets of order 1, 2, and higher order ones, which express complex interactions between terms in specific contexts.

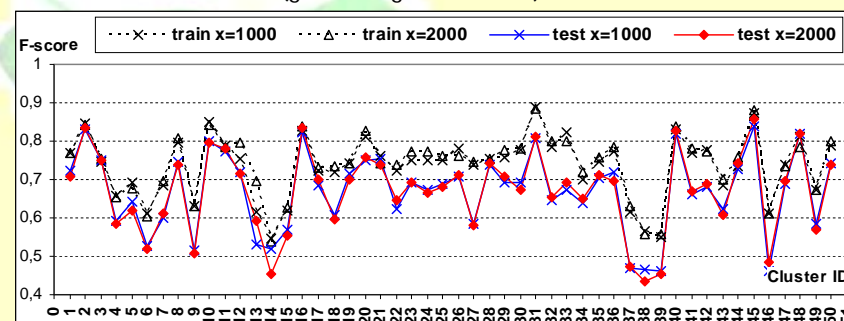
- MIDOVA method (Cadot 2006) for mining ordered lists of informative itemsets specific to each category (train set: 1/10th of the corpus, test set: 9/10th) → ordered lists of simple Boolean queries for identifying the class of any document out of the corpus, and extending this categorization process to other databases (patents, ...).
- Results
 - At the same time as a 100% intrinsically exact reconstitution rate results from using the whole 237,000 terms, a maximum 68% rate results from using about x = 2000 k-itemsets (and then decreases), with a statistically-controlled generalization ability:



- 68% is a mean value, embedding many discrepancies:

>80%: clusters N°45 (LOD/ Linkage_analysis/ Polymorphism) and N°16 (Spectrometry/ Proteomics)

<50%: clusters N°14 (Locus/ Microsatellite_Locus/ Polymorphism) and N° 38 (genome/ genome_size)



- example of class description: the 6 first k-itemsets of cluster 10:

comparative_genomic_hybridization, hybridization AND tumor AND genomic, hybridization AND tumor, CGH AND comparative_genomic_hybridization, losses AND genomic, tumor AND genomic.

As may be observed *comparative_genomic_hybridization* and *CGH* acronym appear together, integrally or partly, at the top of the list.

References

- Bassecoulard E., Lelu A. and Zitt M. (2007). Mapping nanosciences by citation flows: a preliminary analysis, *Scientometrics*, vol 70, n°3, pp. 859-880.
- Cadot M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Doctoral dissertation, University of Franche-Comté, France.
- Laurens P., Zitt M. and Bassecoulard E. (to appear), “Delineation of the genomics field by hybrid citation-lexical methods: interaction with experts and validation process”, *Scientometrics*.
- Zitt M., Ramanana-Rahary S. and Bassecoulard E. (2006). Delineating complex scientific fields by a hybrid lexical-citation method: an application to nanosciences, *Information Processing and Management* Vol. 42-6, pp. 1513-1531.